Rojas-Berscia, L. M., & Roberts, S. (2019). Exploring the history of pronouns in South America with computer-assisted methods. *Journal of Language Evolution*, *4*(3), [lzz006]. https://doi.org/10.1093/jole/lzz006

Journal of Language Evolution, 092019, 1–21
doi: 10.1093/jole/lzz006
Research article

Exploring the history of pronouns in South America with computer-assisted methods

Luis Miguel Rojas-Berscia *,†,‡,§ **and Sean Roberts** **

†University of Queensland—School of Languages and Cultures, St Lucia 4072, Gordon Greenwood Building, QLD, Australia, ‡Max Planck Institute for Psycholinguistics—Department for Language and Cognition, Nijmegen 6525XD, The Netherlands, §Radboud Universiteit Nijmegen—Centre for Language Studies, Nijmegen 6525XZ, The Netherlands and **Department of Anthropology and Archaeology, University of Bristol, 43 Woodland Road, Clifton, Bristol, UK

*Corresponding author: l.rojasberscia@uq.edu.au

Abstract

Pronouns as a diagnostic feature of language relatedness have been widely explored in historical and comparative linguistics. In this article, we focus on South American pronouns, as a potential example of items with their own history passing between the boundaries of language families, what has been dubbed in the literature as 'historical markers'. Historical markers are not a direct diagnostic of genealogical relatedness among languages, but account for phenomena beyond the grasp of the historical comparative method. Relatedness between pronoun systems can thus serve as suggestions for closer studies of genealogical relationships. How can we use computational methods to help us with this process? We collected pronouns for 121 South American languages, grouped them into classes and aligned the phonemes within each class (assisted by automatic methods). We then used Bayesian phylogenetic tree inference to model the birth and death of individual phonemes within cognate sets, rather than the typical practice of modelling whole cognate sets. The reliability of the splits found in our analysis was low above the level of language family, and validation on alternative data suggested that the analysis cannot be used to infer general genealogical relatedness among languages. However, many results aligned with existing theories, and the analysis as a whole provided a useful starting point for future analyses of historical relationships between the languages of South America. We show that using automated methods with evolutionary principles can support progress in historical linguistics research.

Key words: Bayesian phylogenetics; Amerindian linguistics, historical linguistics, pronouns, micro-evolution

1. Introduction

Pronouns have caught the attention of several scholars through history. The father of experimental psychology, Wilhelm Wundt, was one of the first to be interested in the consonants one finds in the pronouns of some languages of the world. After surveying a number of pronominal systems in the languages of Europe and Asia, he realised that many of them deployed a nasal bilabial [m] or a nasal alveolar [n], and concluded:

At last, analogous sound gradations seem to occur pervasively with personal pronouns. This case is also attributable to spatial distance-differences. However, in some cases, another reason might be involved, which provides

the sound metaphors with their peculiar character. Strikingly, the 'I' frequently shows resonance sounds, namely the labial resonance tone m, *in otherwise completely foreign languages. That shows that the natural man, following widespread animistic ideas, transfer his* ego *to his inner body. Thus, the association between the sound articulated with the closed lips and the inner self may be perceived as a natural sound metaphor for the* ego.[1] *(Wundt 1904: 344–5)*

These trends had already been noticed in Amerindian linguistics before. A decade and a half prior to the publication of *Völkerpsychologie* by Wundt, the American archaeologist and ethnologist Daniel Brinton made similar claims with regard to pronouns in America:

[. . .]the N sound expresses the notion of ego, *of myselfness, in a great many tongues, far apart geographically and linguistically. It is the sound at the basis of the personal pronoun of the first person and of the words for* man *in numerous dialects in North and South America. Again, the K sound is almost as widely associated with the ideas of* otherness, *and is at the base of the personal pronoun of the second person singular and of the expressions for superhuman personalities, the divine existences. It is essentially demonstrative in its power. (Brinton 1888: 6–7)*

Explanations beyond sound-symbolism abounded in Amerindian linguistics. One of the most famous paradigms under discussion was the *n: m* pronominal paradigm. For scholars such as Boas (1917) its existence could be explained psychologically. Kroeber (1913) argued that the *n: m* paradigm could be explained in terms of 'territorial continuity of characteristics'. Sapir (1929), Swadesh (1954), Greenberg (1960, 1987), and Ruhlen (1987) followed in claiming that the *n: m* paradigm could not be explained by mere chance. Its existence had to be explained in terms of an undeniable genealogical relatedness (see Campbell 1994 for a detailed historical survey on the topic). This would later on be one of the backbones of Greenberg's Amerind hypothesis (Greenberg 1987).

The latter view was severely criticised by Campbell (1994). The author claimed that the Amerind *n: m* paradigm was largely overstated. This pattern could be explained by:

1. the pervasiveness of nasals in grammatical morphemes, particularly in pronominal markers,[2]
2. the common occurrence of nasals in grammatical morphemes, given their perceptual salience (Maddieson 1984: 70, also cited in; Campbell 1994: 4),
3. contact phenomena, since pronouns can also be borrowed, and
4. child language.

That same decade, Nichols and Peterson (1996) argued that the geographical distribution of the *n: m* pronominal paradigm cast doubt on the Amerind proposal. Based on a sample of 173 languages covering most of the world, the authors showed that the distribution of the *n: m* pronominal paradigm was not exclusive to the whole New World, but was a western American phenomenon. In addition, the paradigm could also be found in Melanesia. Finally, the authors suggested that there was enough evidence to postulate a Pacific Rim historical marker. Historical markers are not diagnostic of genealogical relatedness among languages in the traditional family tree model fashion, but account for phenomena beyond the grasp of the historical comparative method (Nichols and Peterson 1996: 359).[3]

Today, in general linguistics, there is no real consensus regarding the reliability of pronouns as indicators of genealogical relatedness. Dixon (1997: 22), for example, in the same vein as Greenberg and Ruhlen, claimed that pronouns are less likely to be borrowed, and would therefore be 'the surest indicators of genetic relationship'. Matras (2007: 53), from a cross-linguistic perspective, claims that pronominal forms are borrowed, but not as wholesale structural sets, but depending on their functionality. The author resorts to social forces that explain why, for example, Imbabura Quechua developed a special formal second person pronoun *kikin* (from the Quechua reflexive *kikin*) on the basis of contact with Ecuadorean Spanish and the need for a Quechua version of the Spanish *usted*. However, the author also mentions that 'pronouns may be borrowable in principle, [but] show very low borrowability' (Matras 2009: 208). Borrowing as a tendency of pronominal systems has been reported for Pirahã in Thomason and Everett (2001). The authors argue that the pronouns of Pirahã were borrowed from Nheengatu and Tenharim, two Tupian languages known to have been in contact with Pirahã, and then adjusted based on the phonology of the language (2001: 310). The authors, based on this example, conclude that pronouns are not reliable when addressing questions of genealogical relatedness, given their level of borrowability and attitudes behind this process, which are unlikely to be retrievable and understandable from distant contact situations. Other authors remain agnostic. Sasse (2015: 197) argues that pronouns 'are among the most stable elements of basic vocabulary'. In addition, the author adds that pronouns as syntactic categories to account for genealogical relatedness

are reliable for some cases, such as for the reconstruction of Indo-European or Afroasiatic. However, he also emphasises that this seems not to be the case for all the languages of the world. In many cases, similarities across pronominal systems are just due to a tendency for pronominal forms to be simplified and resort to the same set of phonemes, such as *m*, *k*, or *s*. Moreover, the author, in the same vein as Matras (2007), points to the importance of social structure when it comes to the development/diffusion of pronominal systems, that is innovations or borrowings in a system may be the consequence of complex social forces in specific speech communities.

In this article, we do not take part in the *n: m* pronominal paradigm debate and neither are we attributing sound-symbolism properties to pronouns. We do admit, however, that pronouns cannot be taken as direct evidence for genealogical relatedness. Instead of assuming them to indicate the history of the languages they belong to, we prefer to focus on them as items having their own history. Therefore, we do focus on historical markers reflected in South American pronouns, following Nichols and Peterson (1996). Subsystems within lects[4] (cf. Bailey 1973), such as pronouns or pronominal systems, can have their own history. Mufwene, in this regard (2005: 33) suggests,

Although some ethnographic considerations suggest that selection also applies at the level of languages, when speakers target primarily features of a particular language over those of others, what we know about language mixing and the development of creoles suggests otherwise. Languages are selected indirectly through the fact that their features (sounds, words, combinatory rules, and particular ways of packaging meanings) wind up constituting the majority of those selected from the combined feature pool of the language varieties in contact.

Pronouns or pronominal systems, as such, provided they have been selected and become part of a given lect recursively from generation to generation, can help us understand the linguistic history of a population.

South America was populated less than 20,000 years ago. Yet, its linguistic diversity is striking (q.v. Nettle 1999a; Muysken and O'Connor 2014), and in most of the cases not easy to explain in terms of the traditional historical comparative method. A *micro* or item-based approach (q.v. Nettle 1999b; Enfield 2014) may prove more useful. By now, there is a large amount of data available on pronoun systems across many language families in South America. With the development of new computational tools for assessing large-scale data, we see an opportunity to take a fresh look at pronoun systems in this area. We collect and transcribe pronoun systems for 121 languages from 35 South American language families and isolates. With this amount of data, how can we begin the task of spotting patterns and organising the data for future identification of historical markers? How can we avoid cherry-picking individual features or instances that seem appealing, and instead take a broader view of the available data? We take a computational approach to this problem, using recent advances in computer-assisted historical linguistics. There are now many tools for tackling problems such as cognate detection, sequence alignment, and phylogenetic tree inference (see, e.g., List 2019; List et al. 2018; Bowern 2018). However, these mostly rely on having lexical data for many concepts, which might have different histories in South America. In this article, we attempt to apply these methods to a limited number of pronoun concepts by modelling the evolution of individual segments within cognate sets, rather than whole cognate sets. This method would be valuable because it allows the historical inference process to extract more information from the same data. However, it also poses some challenges, such as sequence alignment. The aim is to assess whether this approach is practical, and to produce a useful set of suggestions about how pronoun systems might be related for use in future studies.

The article is structured as follows. Section 2 introduces the use of phoneme-level data in phylogenetic reconstructions. Section 3 presents the methods used for data collection and coding, as well as the details concerning the particular Bayesian phylogenetic analysis we carried out. Section 4 presents the results of the analysis. In Section 5 we discuss possible interpretations of the results, which are still preliminary, given the novelty of the analysis. Finally, we conclude in Section 6 with a to-do list of possible future avenues in the study of the history of South American languages.

2. Using phoneme-level data in phylogenetic reconstructions

The standard method for automatically generating phylogenetic trees in linguistics is to use Bayesian phylogenetic analysis with a Continuous-Time Markov Chain (CTMC) model of cognate birth and death (Bouckaert et al. 2012). Lexical data for many languages is split up into cognate sets (sets of words that are historically related, usually analysed using the comparative method). The CTMC model allows a new cognate to

Langauge	Form	Cognate set
Mangarrai	toko	A
Paiwan	tsuqela	B
Itbayaten	tuqgan	B
Numfor	kor	C

Binary coding

Language	A	B	C
Mangarrai	1	0	0
Paiwan	0	1	0
Itbayaten	0	1	0
Numfor	0	0	1

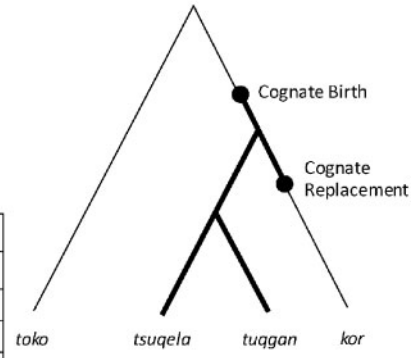

Figure 1. An example of forms grouped into cognate sets from four Austronesian languages (from https://abvd.shh.mpg.de/austronesian/research.php, top left), which are translated into a binary format (bottom left). The right side shows a possible historical tree linking the different forms, with the birth and death of cognate set 2 shown.

appear at a certain point in the tree (birth), to be transmitted along the branches of the tree and then be replaced (death). A given tree structure and a particular set of parameters dictating when the cognate will appear and disappear will fit the data more or less well (Fig. 1).

That is, we can evaluate how likely a tree is to produce the data we see, given the model of birth and death. If there were infinite time to run the analysis, then all possible combinations of tree structures and parameter types could be evaluated and we could select the best one. However, this is not feasible, and so instead a Bayesian Monte Carlo method is used. The space of possible trees and parameters is explored stochastically. The Monte Carlo process gradually converges on a set of trees and parameter values that are a good fit to the data.

This method relies on having multiple cognate sets for multiple concepts. For example, the Austronesian Basic Vocabulary Database includes about 34,000 cognate sets for 400 languages (around 85 cognate sets per language), and the database used in Grollemund et al. (2015) to generate a tree for Bantu languages has 3,859 cognate sets for 424 languages (around 9 cognate sets per language). The different histories of each cognate help narrow down the set of likely tree structures. However, different sets of words have different histories. It should be possible to produce trees based on a particular sub-set of words, such as pronouns. The problem is that there are few concepts and therefore few cognate sets within such a limited sub-set of concepts.

One solution is to look for structure below the level of the cognate (for a similar approach, see Macklin-Cordes and Round 2015). Individual segments of words change, and this can include historical data, too. That is, instead of modelling the birth and death of cognates, we can model the birth and death of segments in particular (aligned) positions within a cognate. For example, Fig. 2 shows four reflexes of the cognate for 'dog' or 'hound' in Dutch, West Flemish, Limburgish, and English. The forms can be aligned into groups that reflect related segments. For example, the word in Dutch, Limburgish and English begins with a/h/, while the West Flemish word does not. This might indicate that West Flemish is further from the other three languages in the tree: at some point, West Flemish lost the /h/ while the others retained it. Some alignment columns have more than one segment, and these can be split into multiple sites. For example, the last segment is either a /t/, /c/ or /d/. That can be translated into three binary features: '/t/ at last position', '/c/ at last position', and '/d/ at last position'.[5]

Therefore, we can generate a set of binary site data that can be analysed with a CTMC model, but now based on birth and death of segments in particular alignment columns. The procedure is:

1. For each concept, split forms into cognate sets.
2. Within each cognate set, align the segments.
3. For each unique segment type within each alignment column, produce a binary vector which shows whether the language has or does not have the given segment.
4. Combine the vectors into a binary matrix.

In the sections below, we apply this method to pronoun data from South America.

2.1 Limitations of the method

Bayesian phylogenetic methods provide a useful tool for investigating linguistic history, but have many issues. The

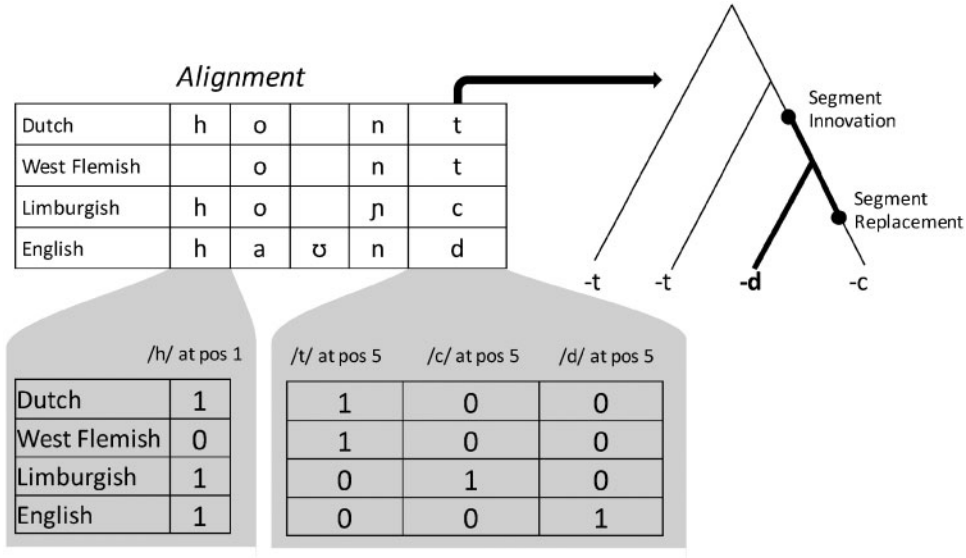

Figure 2. How aligned forms are converted into binary features (left), so that it is possible to model the innovation and replacement of particular phonemes (right). The tree structure is hypothetical, not the suggested best historical analysis for this case.

current approach using phoneme-level data has specific issues which we discuss here. The standard CTMC method assumes that sites are independent, which may not be realistic. For example, in Fig. 2, having a /t/ at the end means that a language does not have a /d/ or a /c/ at the end, and so the three sites are not independent. However, a given language might have multiple segments at a given site if it has multiple forms. Indeed, we find this is often the case in our data (30% of meanings had multiple alternative forms within a language, compared to 3% for comparable data in Pama–Nyungan). That is, strictly speaking, the presence of a particular segment does not imply the absence of all others. Note that this issue also applies for cognate-level analyses, and the same argument applies: a language might have multiple forms for the same meaning which belong to more than one cognate set.

The model also assumes that the pronoun forms within a language evolve independently. That is, it does not take into account the structure of the paradigm. Related to this, contextual sound changes look like two changes in our data, but are really only one change. Failing to account for these issues could make it look like more evolution is happening than in reality, and this could make the branch lengths longer. However, we are not concerned with estimating branch lengths in this study.

Bayesian phylogenetic models are improved by entering prior knowledge about clades. Prior evidence for clades should be independent from the linguistic information, for example historical, archaeological, or genetic evidence. Such evidence is available for relationships between linguistic populations as a whole, but in a micro-evolutionary approach what is needed is evidence that relates to the specific feature. In this case, there is no clear prior evidence that relates directly to pronoun history. Therefore, we did not enter any prior information.

Bayesian phylogenetic methods return binary trees connected by a single root node, which assumes that all taxa in the analysis are related to a single common ancestor. This is not assumed in standard linguistic analyses until a clear genetic family relationship can be identified, which we do not have for the case of South America. While in some distant sense many or all languages in South America may be related, it is not clear that all pronouns are. We proceed with the assumption that they are, with the aim of producing a historical analysis that is most consistent with this assumption. This may be used as a guide for thinking about genealogical relationships and for generating hypotheses, rather than a claim about the true history of the languages.

3.Methods

3.1 Data

Data for four pronoun categories were collected: First person singular, second person singular, third person singular and first person plural either inclusive (sometimes called the 'fourth person' in studies of Andean South American languages[6]) or exclusive. Pronoun data for fifty-nine languages were obtained by the first

Figure 3. Locations of languages in the data.

author.[7] Other sources provided a further sixty-four languages from South America (fifty-two from the Intercontinental Dictionary Series, Key and Comrie (2015); four from the World Loanword Database (Haspelmath and Tadmor 2009), and eight from Birchall, Dunn, and Greenhill (2016) for a total of 123 languages from South America with unique glottocodes (Fig. 3).[8] Where sources listed multiple forms, all were included in separate entries. This resulted in 1,208 forms. The segment inventories were simplified and unified (mainly changes for UTF-8 characters).

All data, processing scripts and analysis scripts are available online: https://bit.ly/2XtdLx4

3.2 Cognate coding and alignment

In order to compare the forms, they needed to be split into cognate categories and then the segments needed to be aligned. This task proceeded in two stages. First, automatic coding and alignment provided a starting point, and then the results were manually edited by the first author. The program Lingpy (List, Greenhill, and Forkel 2017) was used to propose cognate classes and create the automatic alignments (using the 'lexstat' method for clustering, threshold = 0.7, and the 'library' method for alignment, with a custom sound class file). The automatic alignment suggested 248 cognate sets. These alignments were later manually recoded into 37 *pseudo-cognate* sets (changes included both changes to cognate set assignment and in the alignment). The criteria for this process was the following:

- We re-classified the automatic alignments in terms of *pseudo-cognates*. Pseudo-cognates in this sense are aligned words at the segment level whose segments were 'plausibly' historically related. This is based on common sound changes reported in historical linguistics literature. For example, it is common for unvoiced obstruents to become voiced (p, t, k > b, d, g) (q.v. Cerrón-Palomino 2003: 170 for such a process in Quechuan), for postvelar obstruents to become velar (i.e. Proto Quechua *qam* > in contemporary Northern Quechua *kam* (Cerrón-Palomino 2003: 154)), for segments such as /w/ and /j/ to be interchangeable.
- We also used paradigm similarity for some judgements, that is if some sets of words are almost identical (e.g. Muniche pronouns and Arawak pronouns), they were regrouped into a pseudo-cognate set.
- For some words, it was judged that different parts had different histories. Because the analysis is at the level of the segment, these words were split in two and the different sections assigned to the relevant cognate sets. While the differences between the automatic coding and the manual coding seem large, the task was greatly simplified by the automatic coding. The final codings and alignments are available in Supplementary Materials.

The number of edits made to the automatic coding were substantial. For example, about 46% of the segment alignments were altered. The human and automatic coding agreed for 95% of word pairs, which seems high, but simply reflects the small number of cognate pairs compared to non-cognate pairs (randomly permuting the original cognate coding also produces an agreement of 95% on average). Using the B-Cubed measure (see List, Greenhill and Gray 2017) reveals a high precision (0.90), but a low recall (0.22, F-score = 0.36). That is, the human mostly agreed with the pairs that the automatic coding identified as cognate, but also found many more pairs of cognates.

These data were then transformed into a binary feature matrix. Each cognate set within each concept contains multiple forms all of which are aligned into corresponding columns. Each column has potentially

several phoneme types within it. These phoneme types were used as the features (see the explanation above). Where a language had multiple forms within a concept and cognate class, only the longest form was used. Across all languages, there were 1,002 sites (alignment columns). For the Bayesian phylogenetic analysis, small cognate sets with fewer than three forms were removed. The final data included 987 sites in 35 cognate sets for 121 languages (about 8 sites per language; first person: 157 sites in 6 sets; Second person: 173 sites in 7 sets; Third person: 272 sites in 8 sets; Fourth Person: 385 sites in 15 sets). This is still an order of magnitude less data than for studies such as Bouckaert et al. (2012), but similar to the amount of data in Grollemund et al. (2015).

3.3 Bayesian phylogenetic analysis

The binary alignment data was transformed into a NEXUS file format for the Bayesian phylogenetic analysis, placing each cognate set within each concept into a different partition, and adding a contiguity correction for each partition. This partitioning ensures that comparisons are only made within each cognate class (so that, e.g. not all possible phonemes are coded for each cognate set, only the ones that occur for words in that cognate set). The program BEAUti (Drummond et al. 2012) was used to create an input file for the analysis, specifying a single tree model and clock model. We used a Continuous-Time Markov Chain model (Bouckaert et al. 2012) with a relaxed log normal clock model in BEAST (Drummond et al. 2006; Suchard et al. 2018). No priors on the topology of the tree were included. Since we were not aiming to estimate dates, these were not calibrated.

Other models were also run, including a strict clock model, and analyses of the data in a single partition, or also including the cognate codings in the binary analyses files. According to the model fit analyses, the relaxed clock model using only the alignment features yielded the best absolute fit.

The analysis was run for 2,000,000,000 generations. Autocorrelation and convergence checks were carried out using Tracer v1.5 (Rambaut and Drummond 2012). The convergence trace was unstable for much of the run, so the first 75% of runs were treated as a burnin. A total of 74% of parameters had effective sample sizes (ESS) of more than 2,000 (average = 5,910). All parameters had ESS of more than 200, except for 7 parameters (4% of parameters, note that this is better than for the Pama–Nyungan data, see below). An additional run of the model was done, adjusting operator weights to focus on obtaining estimates for these weaker estimates, and another excluding the data associated with these parameters, but this lead to further problems with estimations. The best results were still with the original analysis.

We also ran models with priors on the structures of the trees, reflecting known language families. Results for strict monophyletic priors (forcing languages to be strictly within their language family with no outside languages) lead to lower likelihood than for a non-monophyletic prior (monophyletic mean log likelihood = −13638, sd = 14.6; non-monophyletic mean log likelihood = −13042, sd = 14.8). This was also the case in the Pama–Nyungan analysis. Additionally, the non-monophyletic results were almost identical to the original analysis (mean log likelihood = −13042, sd = 14.7). Following Occam's razor, we have chosen to remain with the original analysis (without priors).

We also fit the data with a stochastic dollo model rather than the CTMC model. It converged well (even trace, all ESS > 2,000), but the trees failed to identify many language family clades and agreement between trees sampled from the posterior was low. Comparing the two models with an AICM test (Baele et al. 2012) showed that the CTMC model provided a substantially better fit (AICM for Stochastic Dollo model = 1,36,073; improvement for CTMC model = 26,533; improvement by over 1,000 units).

A single tree from the distribution may be a poor summary of the analysis. Therefore, 10,000 trees were sampled from the posterior, spacing samples evenly to avoid autocorrelation (at least 100,000 trees between samples). TreeAnnotator (Rambaut and Drummond 2017) was used to produce a maximum clade credibility tree (MCCT): the single and most representative tree drawn from the 10,000 tree sample.

4. Results

4.1 Results for Pama–Nyungan

In order to test the validity of the methods, we applied them to data on Pama–Nyungan pronouns. The CHIRILA database (Bowern 2016) includes cognate coding (by experts) for many concepts in Pama–Nyungan languages and this data was used to infer a phylogenetic tree (Bouckaert, Bowern, and Atkinson 2018). Pama–Nyungan happens to have pronoun data for a similar number of languages, a similar geographic range and a similar time-depth to the data we are considering, and has a similar diversity in pronoun forms (unlike, e.g. Indo-European). We obtained the pronouns from 185 Pama–Nyungan languages (thirty-nine cognate sets), aligned the phoneme sequences using Lingpy,

converted the alignments to binary features and inferred historical relations using Bayesian phylogenetic tree inference (using the same methods as above, see the Supplementary Materials for details). We also ran the automatic cognate coding procedure on the data and compared it to the 'gold standard' cognate coding. The aim was not to assess the validity of the previous analyses of Pama–Nyungan history, but to gain an idea of the performance of our method in the ideal case where there are known relationships between all the languages considered and where the cognate coding is based on standard (non-automatic) historical linguistic methods.

From this process, we obtained several insights. First, cognate coding in Lingpy for small numbers of concepts has adequate precision, but tends to split cognate sets more than for larger datasets. This is what we suspected of the results of the automatic coding for the South American data, and indeed most of the manual corrections involved lumping the automatic cognates together. Secondly, the convergence of the Bayesian phylogenetic method was poorer for the Pama–Nyungan data than for the South American data (in terms of parameter ESS). The original inference by Bouckaert et al. used monophyletic priors on the sub-families, but for pronouns the likelihood was better without priors than with monophyletic priors. This might reflect pronouns being borrowed between sub-families. Finally, the results for Pama–Nyungan failed to recover many of the sub-families or the internal structure of many sub-families. We therefore conclude that the phoneme-level method for restricted concept sets *cannot* be used to reliably infer the general history of whole languages. However, it may still be useful for rapidly producing a first approximation of the history of particular sets of words, and for generating hypotheses that can help historical linguists in future studies.

4.2 Results for South America

Densitree (Bouckaert and Heled 2014) was used to visualise the distribution of 10,000 trees drawn from the posterior. Figure 4 shows the densitree: each tree in the sample is drawn on top of each other so that overlapping lines show more strongly. It is clear that there is some agreement near the tips of the tree, but that the agreement near the root is very poor.

The MCCT is shown in Fig. 5. The numbers at each node in the MCCT indicate the proportion of trees in the posterior sample that contain the same split (higher numbers indicate greater support for the split). It is clear that several branches have poor support: the sample of trees do not agree on the split, suggesting that it is not reliable. We note that the agreement for the South American data was generally better than for the Pama–Nyungan data.

Some expected splits are observed. For example, the tree splits languages into language families reasonably well. This can be tested by comparing the tree produced from pronouns illustrated in Fig. 5 to the Glottolog classification (Hammarström, Forkel, and Haspelmath 2018). For all languages in the data, we joined their Glottolog families together into a single tree (binding at the root, with unclassified languages being attached at the root). Comparing true distances between trees is computationally hard. Instead, the quartet distance measure can be used (Estabrook, McMorris, and Meacham 1985; Pompei, Loreto, and Tria 2011). For all possible quartets of languages, it produces a sub-tree for each of the trees being compared, and then compares the overlap between these sub-trees. This produces a distance measure that assesses the structural distance between trees (it does not take branch lengths into account). To assess the significance of this distance, the same test can be run but where the tips of one of the trees is randomly permuted. Doing this many times gives a distribution of permuted tree distances in order to calculate a z-score (how far the true distance is from the mean permuted distance, in number of standard deviations of the permuted distance) and p value (the proportion of permuted trees that result in a smaller distance). We found that the tree produced from pronouns is significantly similar to the Glottolog classification compared to trees where the tips have been randomly permuted (true quartet distance = 7,934,529, mean distance of 1,000 permutations = 8,234,177, $z = -17.9$, $P < 0.001$). However, the pronoun tree is not significantly closer to the Glottolog tree when permuting languages only within language families (mean permuted distance = 7,942,530, $z = -0.60$, $P = 0.26$). This suggests that the pronoun tree is classifying languages into Glottolog families well, but it does not agree with Glottolog below the level of families. These results are similar for the Pama–Nyungan data ($z = -23.87$, permuting within 'families' $z = -4.62$).

We also compared our results to the phylogenetic tree of Chapacuran languages produced by Birchall, Dunn, and Greenhill (2016) (Fig. 6). Both trees place Kitemoka and Chapakura together (and also Wari' and Oro Win), but the rest of the tree is quite divergent.

The distances in the MCCT were weakly but significantly correlated with geographic distance between languages (geographic coordinates from Glottolog, Mantel test $r = 0.13$, $P < 0.001$).

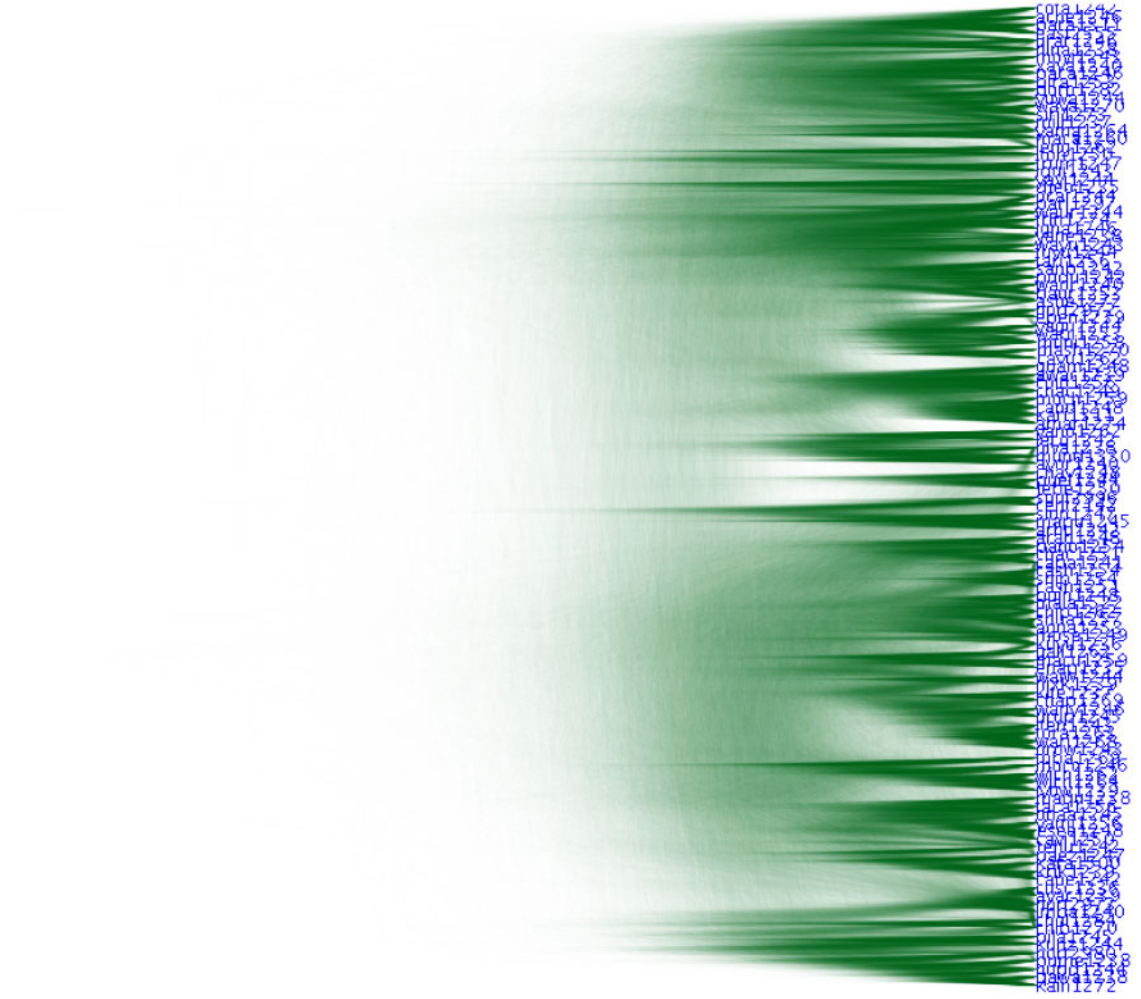

Figure 4. A densitree of the posterior distribution.

4.3 Examples

Parts of the tree nearer the tips do have reasonable support. For example, each node in the sub-tree connecting Maquiritari, Hixkaryana, Waiwai, Galibi Carib, E'napa Woromaipu, and Macuchi pronouns—all pronouns of Carib languages—are observed in over 95% of trees (Table 1).

Another example is that of Eastern Bolivian Guarani, Paraguayan Guarani, Mundurukú, Aché and Urarina pronouns, the first four of which belong to the Tupian language family, and the last one is considered unclassified. Each node in the sub-tree connecting these pronouns is observed in over 75% of the trees (Table 2).

The first example (Table 1) shows the degree of agreement of our analysis based on four concepts with the family level, that is four concepts were enough for our methods to infer these forms were related. The second example (Table 2) shows some noise (see e.g. Urarina pronouns, which are included in this part of the tree and do not show any straightforward similarity with the rest of pronouns), which calls for future refinements in the methodology.

5. Discussion

The reliability of splits in the tree was low for splits above the level of language family and in comparison to several other studies. However, the results are still reasonable considering that it only used lexical data from four concepts and looked at a range of languages with a much deeper time depth than most other phylogenetic studies. In the following sections, we examine some of

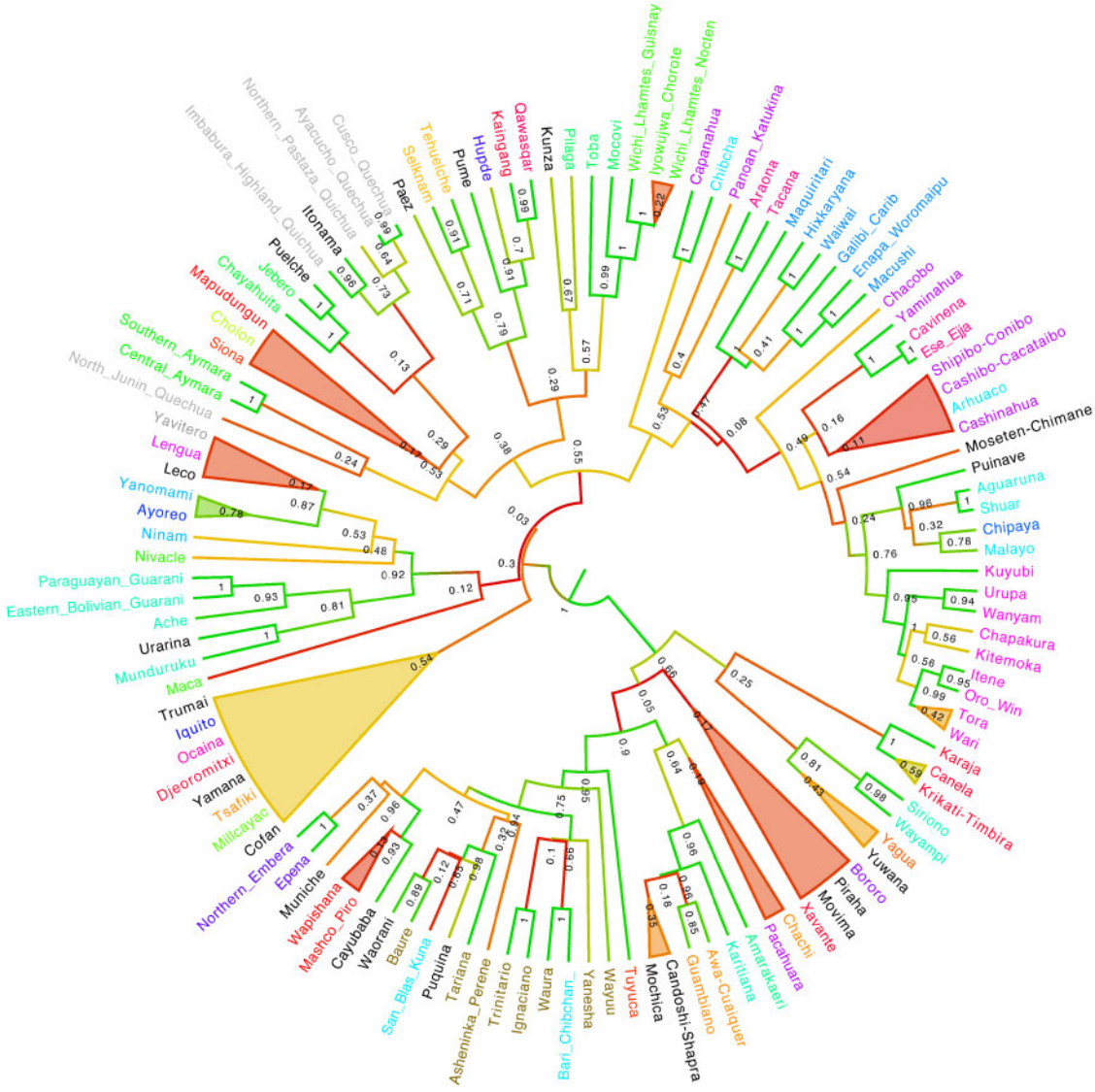

Figure 5. The MCCT. Labels are coloured according to Glottolog language family. Numbers indicate the proportion of trees in the sample that include the given split and the branches are coloured according to the support (green = good, red = poor). All nodes in collapsed sections of the tree have posterior probabilities of less than 80%.

the results, which have sometimes been addressed in the literature and could eventually shed light on the linguistic past of the continent.

5.1 The Kawapanan–Puelche hypothesis and beyond

One of the sub-clades that shows a strong support is the one formed by Kawapanan languages (Chayahuita and Jebero) and Puelche, separated by a distance of ca. 5,400 km (Fig. 7).

Kawapanan and Puelche pronominal systems display a striking similarity (Table 3).

Even third person pronouns, which are sometimes prone to borrowing or taken from the demonstratives/deictics system in most languages, are also strikingly similar. This seems to have been observed previously by Kaufman (1990, 1994). The author argued for the plausibility of existence of a Macro-Andean family, which includes Kawapanan, Puelche, Urarina, Warpean, Jivaroan, Bora-Witoto, Zaparoan, Taushiro, Omurano, and Waorani. However, given the lack of actual

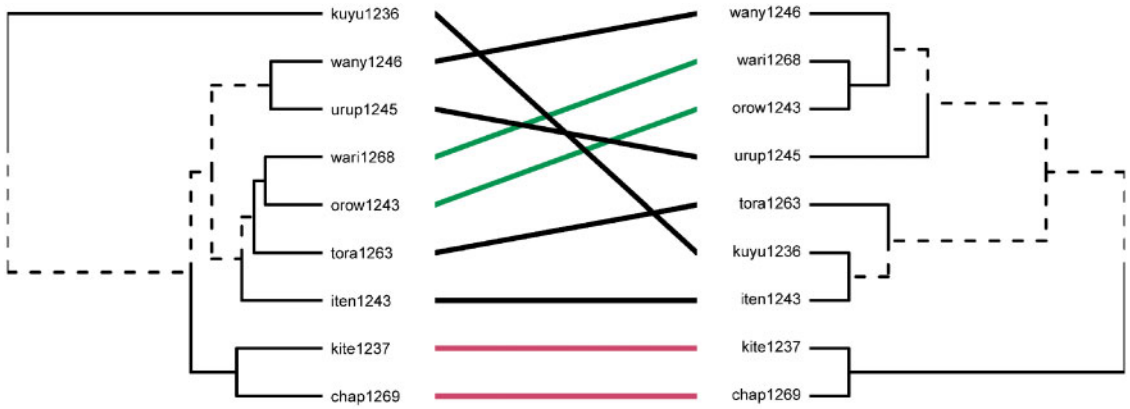

Figure 6. A tanglegram comparison between the tree for Chapacuran languages from this study (left) and from Birchall, Dunn, and Greenhill (2016) (right).

Table 1. Maquiritari, Hixkaryana, Waiwai, Galibi, Carib, E'napa Woromaipu, and Macuchi pronouns.

	I	you (singular)	he/she/it	We
Maquiritari	iwi	əmədə	tiwi	kiwi/nna
Hixkaryana	uro	omoro	noro	kɨwro
Waiwai	owi	amoro	ero/noro	amna/kiiwi/kiwyam
Galibi_carib		amolo	moko/molo	kiko/nana
Enapa_Woromaipu		amən	mək/ mukuh	ana/ we ʔnəkon/yutokon
Macuchi	uuri	amiri	miikiri	anna/uuri ʔnikon

Table 2. Eastern Bolivian Guarani, Paraguayan Guarani, Mundurukú, Aché, and Urarina pronouns.

I		you (singular)	he/she/it	we
Aché	čo	ǰe	go	naǰe; ore
E. Bolivian Guaraní	če; se	nde	hae	yande; ore
Paraguayan Guaraní	če	nde	haʔ e	haʔe; haʔe čupe; ande; ore
Munduruku	õn	ẽn; e	išẽ; ite; ibo; iǰa; iǰop; io	wy-jĩ; oče – jĩ
Urarina	kanu	i:	aka	kana

argumentation with proper formal correspondences, this classification remains anecdotal.

It is still difficult to claim that there is a clear genealogical relatedness between Kawapanan and Puelche, given the lack of proper formal correspondences, but evidence for a non-chance/historical connection is compelling. Even the comparison of some verbal morphology, such as subject or object indexation, reveal considerable affinity (Tables 4 and 5):

In addition, some pronominal number markers such as 3aug *-na* –for both Kawapanan and Puelche –, and dual *-npu*/*-w(p)*, as well as the deictic marker *–su*$^{'9}$/*ʂa*—in Kawapanan and Puelche, respectively—show a strong resemblance (Table 6). Below, we also provide some affinities in the core lexicon of both language groups:

Moreover, one of the most remarkable systematic affinities between the Puelche and Kawapanan pronominal systems is the presence of an initial *k-*. According to Viegas Barros (2017) the *k-* present in Puelche pronouns would be a pronominal base or pronominal marker. From an internal-reconstruction perspective, this could have also been the case for Kawapanan.

The languages geographically located between Kawapanan languages and Puelche seem to have retained some of the formal features of this pronominal system (Table 7):

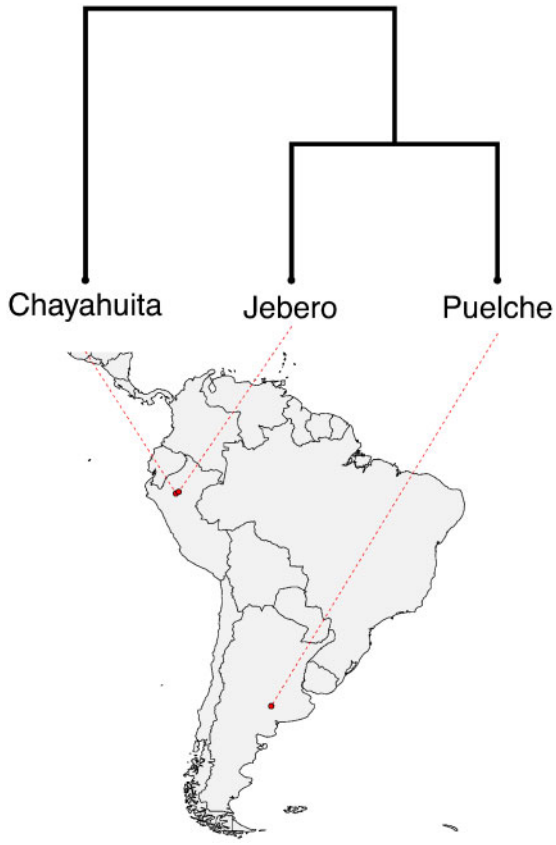

Figure 7. The Kawapanan–Puelche pronominal clade.

Table 3. The Kawapanan and Puelche pronominal systems. The table also includes Proto-Kawapanan (Valenzuela Bismarck 2011; Rojas-Berscia and Nikulin 2016) for comparative purposes.

	*Proto-Kawapanan	Shawi	Shiwilu	Puelche[a]
1	*kwa	ka	kwa	kwa/kja
2	*kɨma[b]	kɨma	kɨnma	kɨmaw
3	*saja?[c]	ina/saja	nana	ʂaʂa
1.pl	*k(i/u)ja	kija	kuða	kɨja/kɨʂa

[a]These forms can be found in Malvestiti and Orden (2014). The forms were standardised to the IPA, on the basis of explanations for letter-sound correspondence provided by the authors and the insights of Casamiquela (1983).

[b]Throughout the article we use /ɨ/ when referring to the Kawapanan phoneme /ɘ/, given that this was the form we used in the original Bayesian phylogenetic analysis.

[c]This form is attested in the earliest forms of Chayahuita, also known as Mayna-Chawi (see Rojas Berscia, 2015).

The languages, the pronominal system of which is presented in the chart, with the exception of Warpean, were clustered together (Fig. 8). Although pronominal information is not enough to claim a precise vertical relatedness between these languages, the similarities between the pronominal systems could account for the existence of a potential historical marker (Nichols and Peterson 1996), which would extend all the way from the south-west of modern Colombia to Tierra el Fuego. A process which could have been fossilised in the pronominal systems of all these languages could be that of the early peopling of the western territories of South America, prior to the sedentarisation and local establishment of certain groups due to farming and agriculture (see Diamond and Bellwood 2003). However, this remains hypothetical. We dubbed the clade in the MCCT including these languages and some others 'Andean clade'. Below, we present its distribution.

Table 4. The Kawapanan and Puelche subject suffixes/prefixes systems, based on Rojas-Berscia and Nikulin (2016) for Proto-Kawapanan (PK) and Viegas Barros (2017) for Puelche.

	PK subject suffixes	Puelche subject prefixes
1	-*wɨ*	*u-*
2	-*n*	*mi-*
1.pl	-*i*	*jɨ-*

Table 5. The Kawapanan object suffix system compared to the Puelche second prefix system.

	PK object suffixes	Puelche second subject prefixes
1	-*ku*	*ku-*
2	-*kɨn*	*kɨmɨ-kum*
1.pl	-*kui*	*kɨi-*

Given that we only relied on pronominal information, a more detailed study that considers other structural features in the languages concerned (see Muysken and O'Connor 2014 for a first survey of structural features in the continent) could certainly improve our understanding of this area and discard some possible confounds in the data.

5.2 The Jivaroan–Uru connection

In the last two decades there have been efforts in the description of the grammars of the Uru languages, from a philological (Cerrón-Palomino 2016) and a descriptive perspective (Cerrón-Palomino 2006; Muysken and Hannss 2006), as well as in the analysis of the history behind the population displacements of these people (Barbieri et al. 2011).

Jivaroan or Chicham languages are somewhat in the same situation. Many varieties have been carefully

Table 6. Proto-Kawapanan and Puelche lexical affinities.

Meaning	Proto-Kawapanan (adapted from Valenzuela Bismarck 2011; Rojas-Berscia and Nikulin 2016)	Puelche (Malvestitti and Orden 2014)
water	**jɨk*	*jagup*
drink	**uk*	*gu-*
eat	**kap-*	*kn-*
egg	**kaju*	*gigɨ*
man	**kɨma*	*gina*
blood	**wɨla-yɨk/wiNa-yɨk*	*ginaw*
liver	**kan*	*gin*
face	Sha. *yapi-ra*	*y-apɨk*

Table 7. A comparison of the languages between the Kawapanan–Puelche space. Forms that show some resemblance appear in bold.

	PK (Valenzuela Bismarck 2011)	Puelche (adapted from Malvestitti and Orden 2014)	Proto-Chonan (Viegas Barros 2017)	Proto-Quechua (Cerrón-Palomino 1987, 2003)	Proto-Aymara (adapted from Cerrón-Palomino 2000)	Warpean (Tornello et al. 2011)	Kunza (Peyró García 2005)	Cholón (Alexander-Bakkerus 2005)
1	**kwa*	***kwa/kja***	**ja:*	**ja*	**na-ja*	***ku***	***aka***	***ok***
2	**kɨma*	***kɨmaw***	**kma:*	**qam*	**huma*	***ka***	***cema***	*mi*
3	**saja*	***ṣaṣa/jaja***	**ta:*	**paj*	**up^{b}a*	***eguj/pi/wen***	***aja***	*pi*
1.pl	**k(i/u)ja*	***kija/kiṣa***	**sa:/wékwa:*	-	**hiwa-sa*	***kuchu***	***kuna***	***sa***

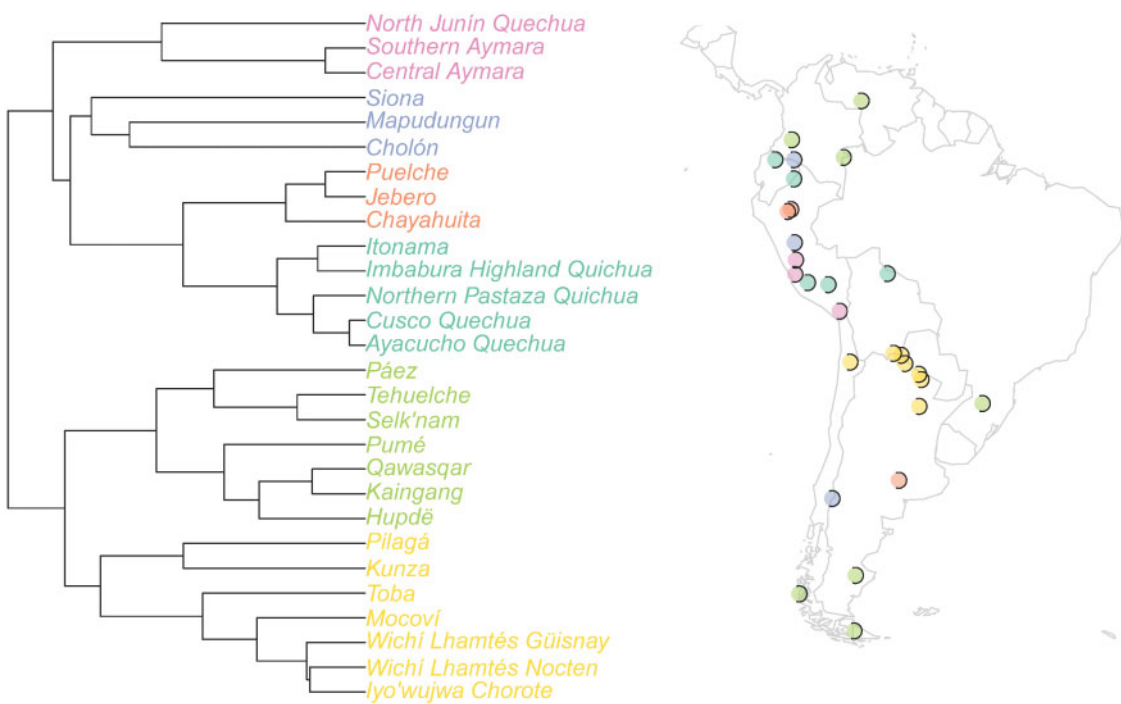

Figure 8. Geographical distribution of the clade.

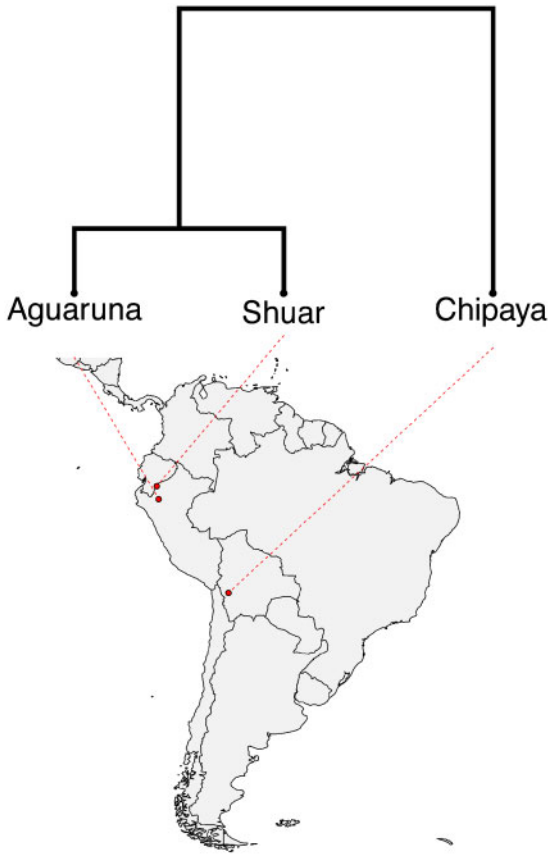

Figure 9. Geographical distribution of the Uru-Jivaroan pronominal clade.

described in the past two decades (Saad 2014; Peña 2016; Overall 2017), and work on ancient DNA with regard to the genealogical affinities between this group and that of the Lamista of San Martín in Perú has shed light on the possible population displacements of the speakers of these languages (Sandoval et al. 2016).

In our analysis, both Jivaroan (Shuar and Aguaruna) and Uru (Chipaya) were clustered together (Fig. 9).[10] The pronominal systems of both groups also show some striking similarities (Table 8):[11]

The affinities between the Uru languages (Chipaya, Iru-wit'u, and Ts'imu) and Jivaroan languages (Awajún and Shuar) are more prominent in the first, second, and third person singular and the second person plural. In addition, both Jivaroan and Uru also appear to share several lexical items (Table 9):

We surmise this to be a case to be considered in future comparative studies involving the Jivaroan or Uru language families. Cerrón-Palomino (2016) provides a complete survey of previous efforts made by comparative linguists to

Table 8. Uru–Jivaroan pronominal systems.

	Chipaya	Ts'imu	Iru-wit'u	Awajún	Shuar
1	*weril*	*witr*	*wiril*	*wi*	*wi*
2	*am*	*Ama*	*am*	*amɨ*	*amɨ*
3	*nii*	*ni*	*ni*	*nu*	*nĩi/au*
4	*utrum*			*hutii*	*ii*
2 pl.	*antsoxo*			*atumɨ*	*atum*

suggest groupings involving Uru and Puquina, and Uru and Mosetean languages. The latter seems to be the only promising avenue in the field. However, no comparative analysis involving geographically very distant languages has been successful. We suggest the Jivaroan–Uru connection to be a promising one from a traditional perspective.

5.3 Puquina, Muniche and Chocoan: the Arawak Matrix

Another grouping which resulted from our analysis and which is worth having a look involves the Puquina (isolate), Muniche (isolate), the Chocoan language family and some Arawak languages (AWK) (Fig. 10). The pronominal systems of all these languages show great affinities (Table 10):

We suggest that the presence of this pattern in non-Arawak languages, such as Northern Embera, Epena, Muniche, and Puquina, is a trace of contact with Arawak speakers. This was also addressed from a structural perspective as the Arawak matrix (Eriksen and Danielsen 2014).

It has been well documented that Arawak itinerants used the major Amazonian rivers to reach places as far as Southwestern Amazonia (Fig. 11). These migrations are not only of archaeological and ethnohistorical interest, but also show a clear linguistic impact. Ritualistic lexicon is the most common fossil of these migrations in the vocabularies of many Amazonian languages. Nevertheless, grammar was influenced in many important ways as well (Rojas-Berscia, 2019). The Arawak pronominal system diffusion could point to this.

Similarities between Muniche and Arawak (Gibson 1996), as well as Puquina and Arawak (Kaufman 1990; Adelaar and Muysken 2004; Jolkesky 2016) have been observed in the literature. Here we confirm these observations from a phylogenetic perspective at the pronominal level, but also add Chocoan languages to the discussion. Although we found the observations not strong enough for us to claim the existence of a genealogical relationship between Arawak languages and Puquina or Muniche, the similarities throughout the entire pronominal system are not coincidental. Once more, we suggest this is evidence of a historical marker, the origin of which is irretrievable

Table 9. Lexical similarities between Uru and Jivaroan. Possible formal affinities are shown in bold.

	Ts'imu	Iru-wit'u	Chipaya	Awajún[a]
who		***heki***	***hek***	***yáki***
no	***ara***	***ana***	***ana***	***atsá***
leaf	***tuk***	***tuk***	*chañi*	***dúka***
frog			***skarap***	***súakaraip(a)/(i)*** 'type of edible frog that sings during the night'
dog	***paku***	***paku***	***paku***	***páki*** 'wild pig'
louse	***sañis***	***sami***	***sami***	**sámig** 'type of centipede'
meat	***hilli***	***xilli***	***chiswi***	**shikíat** 'steamed meat in *patarashca*'[b]
two	***pati***	*pisk*	*pisk*	**pachi-** 'relative to a group, multitude'
seed	***muxu***	*nonis*	***muxu***	***tuxu***
bark	*cici-k'ispi*	*tsts-kispi/**chaxpi***	***chapi***	***saepé***
head	***paeqe***	*ača*	*ača*	***búuk***
ear	***k'uni***	***kuñi***	***khuñi***	***kuwísh***
eye	*c'uxñi*	*cuki*	*chuki*	***xíi***
teeth	***iske***	***iski***	***isqe***	***ixíg(ka)*** **'teeth root'**
bite	***c'at***	***cati-***	***c'at***	***esát***
kill		*kon-*	*kon-*	*kuwaut* 'cut'?
walk		***okx-***	***oqh-***	***wekagát***
give	***ta-***	***ta-***	***tha:-***	***atákut***
say	***ci-***	***ci-***	***khiy-***	***čičát***
floor		***yeku***	***yuqa***	***nugka***
fire	***ux(i)***	***uxi***	*ux*	***híi***
ash	*killa*	*khilla*	***qhup***	***yuku***
path	*litris*	***hiks/liksi***	***hiks***	***hinta***
red	***par***	*xloki*	*lok*	*kapántu*
name	***tuxya***	***tu:***	***thu:***	***da***, *na*

[a]The information for Awajún is based on Antunce, Jakway and Wipio (1996).
[b]A typical Amazonian type of meat-steaming using *bijao* leaves (*Calathea luthea*).

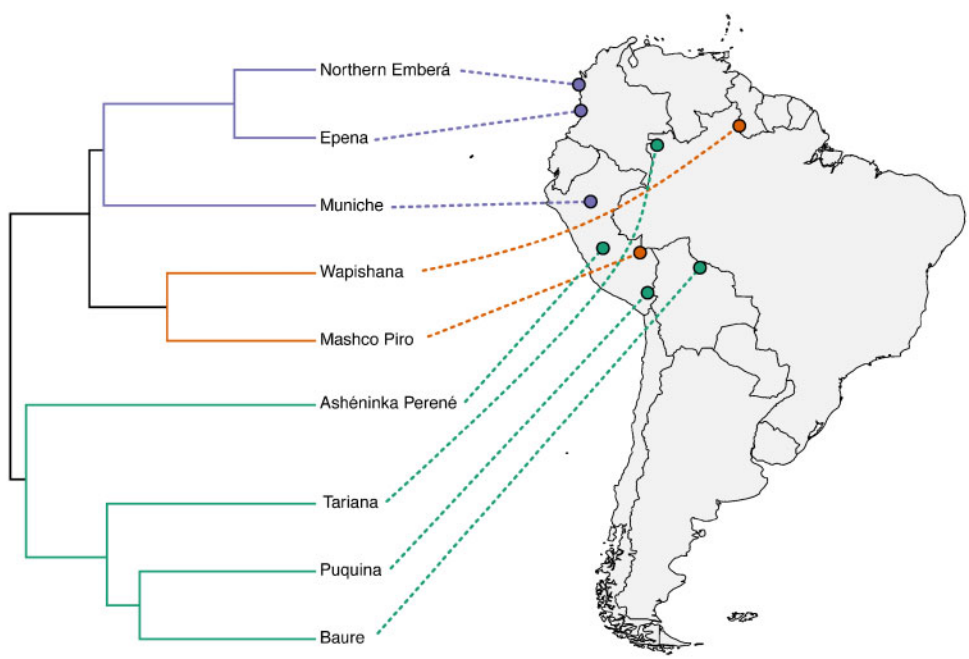

Figure 10. A distribution of the Arawak-type pronominal system.

Table 10. Pronominal systems of Muniche, Northern Embera (Chocoan), Epena (Chocoan), Wapishana (AWK), Mashco Piro (AWK), Tariana (AWK), Puquina, Baure (AWK), and Perené Ashéninka (AWK).

	Muniche	Northern Embera	Epena	Wapishana	Mashco Piro	Tariana	Puquina	Baure	Perené Ashéninka
1	*aʔpa-nü*	*mi*	*mi*	*ugari*	*no*	*nu*	*ni*	*ni/nti*	*no*
2	*aʔpa-pü*	*bi*	*pi*	*pigari*	*pi*	*pi*	*pi*	*piti*	*pi*
3	*aʔpa-ra*	*iru*	*iru*	*uruu*	*wale*	*di*	*tʃ̂u*	*roti*	*i*
4	*aʔpa-wü*	*dai*	*taci*	*wainau*	*wita*	*wa*	*seɲ*	*βiti*	*a*

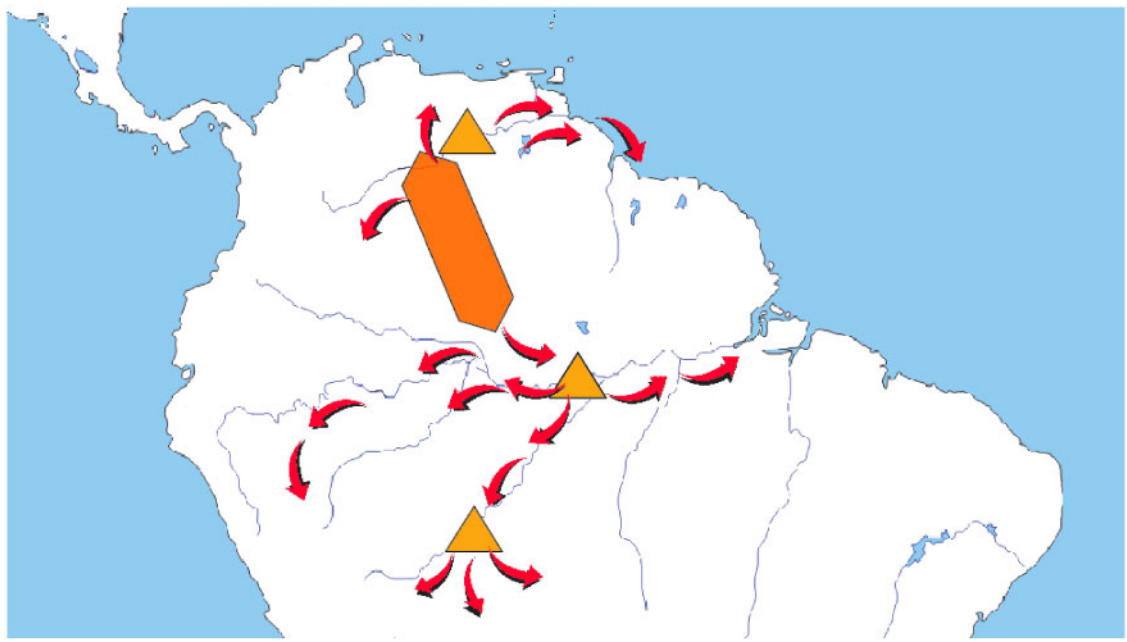

Figure 11. Proposed locations of Proto-Arawakan languages in Northwestern Amazonia (orange hexagon); secondary centres of dispersal in the Middle Orinoco, Central Amazon, and Upper Madeira (yellow triangles); and major routes of expansion (red arrows; adapted from Heckenberger (2002: 116), extracted from Rojas-Berscia and Piepers (in prep.)).

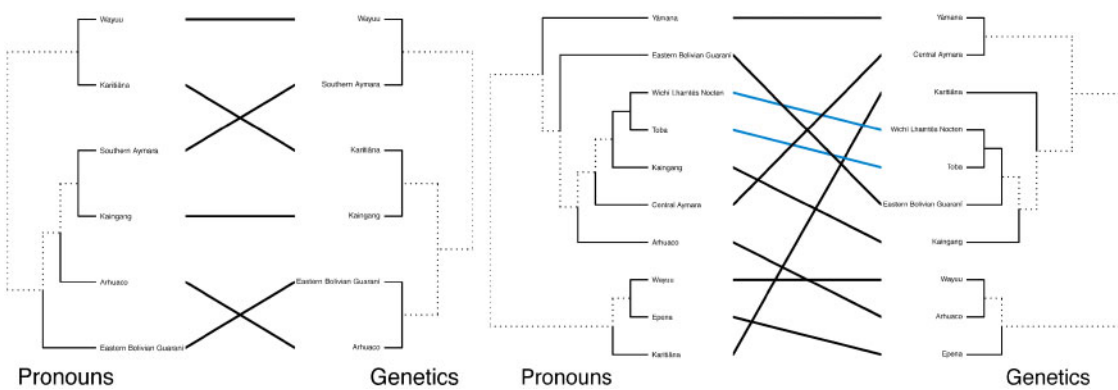

Figure 12. Comparison between the tree derived from pronouns and trees derived from genetic data. Left: Pemberton et al. (2013); right: Reich et al. (2012).

based solely on the comparative method. Therefore, Chocoan, Muniche, Puquina, and Arawak languages may only be historically related at the pronominal level.

6. Concluding discussion

This is just a first step towards more fine-grained studies on historical relations between South American languages from a phoneme-level Bayesian phylogenetic perspective.

Some of the main outcomes from the current analysis were:

1. Well-established language families were recovered by just resorting to pronominal information (see Quechuan, Aymaran, or Carib). Pronominal systems may be a reliable hypothesis generating starting point when it comes to trace back historical (vertical or horizontal) relations. However, this was not the case for the Pama–Nyungan data, and so care is needed in interpreting the history of restricted sets of words as the general history of languages.
2. There is not enough evidence to claim that Puelche and Kawapanan are genealogically related, as was suggested in Kaufman (1994) with his Macro-Andean hypothesis. Not enough shared vocabulary has been found yet. Still, the languages seem to have related pronominal systems. This relationship may reflect a historical marker of the early peopling of western South America, prior to the emergence of farming and agriculture (Diamond and Bellwood 2003), formalised in the presence of a particular k-type pronominal system not only in Kawapanan and Puelche, but also in other Andean and piedmont languages and language families, such as Kunza, Cholón, Quechuan, Aymaran, Warpean, and Chonan.
3. Jivaroan and Uru languages also seem to be historically related at the pronominal level. Although no grouping of these languages has been suggested in the past, there seems to be enough lexical commonalities to carry out deeper comparative studies. Their relationship may well turn out to be genealogical.
4. Puquina, Muniche, and the Chocoan languages have very similar pronominal systems. This seems to be the relic of migrations of Arawak itinerants and the possible scenarios where Arawak vernaculars mixed with local ones. Although hypothetical, this seems to be the most accurate explanation for the diffusion of the Arawak pronominal system into the grammar of these languages. Further lexical and grammatical comparative analysis must be carried out in order to compare the degree to which Arawak-type linguistic elements pervaded the different grammars of Muniche, Puquina, and Chocoan speakers.

The method still needs to be improved and validated on known historical relations. For example, the current analysis treats each segment as categorically different, which can lead to very similar forms looking different. This might be improved by using sound classes instead of segments (we thank an anonymous reviewer for this suggestion). We also note that it would be possible to reconstruct the probability of each segment at each node in the phylogeny, potentially providing a way of suggesting reconstructed forms.

Finally, the use of structural features in the languages of the zone (Muysken and O'Connor 2014) will possibly clear up the picture and allow us to get rid of possible confounds triggered by the use of a unique pattern. Eventually, only a careful interdisciplinary approach which encompasses the study of ancient DNA (Reich et al. 2012; Pemberton et al. 2013) and archaeology will help us support or put into doubt the plausibility of the historical markers we presented and shed light on the linguistic past of South America.[12]

Acknowledgements

L.M. R.-B. would like to acknowledge the financial and academic support of the Language in Interaction Research Consortium in the Netherlands and a Postdoctoral Fellowship granted by the Australian Research Council Centre of Excellence for the Dynamics of Language (CoEDL). S.R. is supported by a Leverhulme early career fellowship (ECF-2016-435). L.M. R.-B. and S.R. thank the anonymous reviewers, Fiona Jordan, and Simon Greenhill for comments and suggestions. L.M. R.-B. would also like to thank the comments and suggestions of Frank Seifart, Stephen Levinson, Pieter Muysken, and Pieter Seuren.

Supplementary data

Supplementary data is available at *Journal of Language Evolution* online.

Notes

1. Translation from the original: Analoge Lautabstufungen scheinen endlich bei dem Personalpronomen weitverbreitet vorzukommen. Dieser Fall ist wohl ebenfalls den räumlichen Entfernungsunterschieden zuzuordnen. Doch dürfte in manchen Fällen noch ein anderes Moment mitwirken, das der Lautmetapher ihren eigenartigen Charakter verleiht. Auffallend häufig kommen nämlich für das 'Ich' die Resonanzlaute, namentlich der labiale Resonanzton *m*, in sonst

gänzlich stammesfremden Sprachen vor. Da schon der Naturmensch nach weitverbreiteten animistischen Vostellungen sein Ich in das Innere seines Körpers verlegt, so mag die Assoziation des bei verschlossenen Lippen vorgebrachten Lautes mit dem eigenen Innern hier als eine natürliche Lautmetapher für das Ich empfunden werden (Wundt 1904: 344–5).

2. The same year, Willerman (1994) conducted a survey with pronoun paradigms of thirty-two typologically diverse languages. Willerman found that nasals were twice as frequent in pronouns. According to this author, nasals seem to be one of the most common segments in pronominal paradigms not necessarily due to sound-symbolism but because of being one of the simplest ones from an articulatory standpoint.
3. This opened a debate in the field (see Campbell 1997 as a response to Nichols and Peterson 1996; and Nichols and Peterson 1998 as a response to Campbell). This debate has been recently reassessed by Zamponi (2017). For Zamponi, '[t]he paradigm observed through the proto-languages of all multi-member families and all isolates of the New World we know, appears essentially as a North American phenomenon that cannot be circumscribed precisely to a single area and is deeply-rooted only in the southern half of the Northwest Coast and in the neighboring California' (2017: 225).
4. A lect is defined as 'a completely non-committal term for any bundling together of linguistic phenomena' (Bailey 1973: 13).
5. In the example above, having a /t/ at the end means that a language does not have a /d/ at the end, and so the three sites are not independent. Independence between sites is an assumption of the Bayesian phylogenetic method. However, strictly speaking, a given language might have multiple segments at a given site if it has multiple forms. Indeed, we find this is often the case in the data. This is also the defense for cognate set coding: a language might have multiple forms for the same meaning which belong to more than one cognate set.
6. We decided to pick up the first four pronoun categories (1s, 2s, 3s, 1pl/4), since these forms commonly appear uninflected in most South American languages. The most dubious choice is 1pl or 4, which is inflected in many cases and uninflected in others (as in Andean languages), and which is sometimes split due to clusivity (inclusive/exclusive). In this case, we picked the form which was less likely inflected from the first singular.
7. The data obtained by the first author was collected in 2014 for the Master of Honours Project Person-marking in the Andes, a historical and comparative perspective. This database has also been made online in our GitHub repository.
8. A reviewer asked why were pronouns used, as opposed to ASJP data. The ASJP has two weaknesses:
   - The set of segments used in the transcriptions is very coarse, with few distinctions in key phonemes for South America (see the 'ASJPcode' section here: https://en.wikipedia.org/wiki/Automated_Similarity_Judgment_Program). Our method depends on phonemic detail.
   - The set of meanings for pronouns only includes 'I', 'you (singular)' and 'we'.
9. *-su'* in deictics is a fossilised form of the contemporary nominaliser *-su'*.
10. This clade also included Malayo, a Chibchan language of Colombia. Although we surmise this as a possible confound, a better sample with more structural features from Malayo and possibly other Chibchan languages could clarify the picture.
11. The phylogenetic analysis only included the 1, 2, 3 and 4 persons of Chipaya, and Awajún and Shuar. In the table we present a more complete table, including data from the Uru languages Ts'imu and Iwu-wit'u, all extracted from Cerrón-Palomino (2016).
12. Currently, there are few alternative data which connect with the current results. The GeLaTo database (http://www.shh.mpg.de/553680/gelato-genes-and-languages-together) is a promising resource that aims to connect data on genetic variation to languages. For example, it links six of the languages in this study to genetic differences in microsatellite markers of their speakers (from Pemberton, DeGiorgio and Rosenberg 2013). Another study (Reich et al. 2012) looked at differences in single nucleotide polymorphisms for American populations and ten languages overlap with ours. These trees are compared with ours in Fig. 12. The similarity is low. This obviously requires much further research.

References

Adelaar, W. F. H., and Muysken, P. C. (2004) *The Languages of the Andes*. Cambridge: Cambridge University Press.

Alexander-Bakkerus, A. (2005) *Eighteenth-Century Cholón*. Utrecht: LOT, Landelijk Onderzoekschool Taalwetenschap.

Antunce, S. A. P., Jakway, M. A., and Wipio, D. G. (1996) *Diccionario aguaruna—castellano, castellano—aguaruna*.

Lima: Ministerio de Educación & Instituto Lingüístico de Verano.

Baele, G. et al. (2012) 'Improving the Accuracy of Demographic and Molecular Clock Model Comparison While Accommodating Phylogenetic Uncertainty', *Molecular Biology and Evolution*, 29/9: 2157–67.

Bailey, C.-J. N. (1973) *Variation and Linguistic Theory*. Arlington, Virginia: Center for Applied Linguistics.

Barbieri, C. et al. (2011) 'Mitochondrial DNA Variability in the Titicaca Basin: Matches and Mismatches with Linguistics and Ethnohistory', *American Journal of Human Biology: The Official Journal of the Human Biology Council*, 23/1: 89–99.

Birchall, J., Dunn, M., and Greenhill, S. J. (2016) 'A Combined Comparative and Phylogenetic Analysis of the Chapacuran Language Family', *International Journal of American Linguistics*, 82/3: 255–84.

Boas, F. (1917) 'Introduction', *International Journal of American Linguistics*, 1: 1–8.

Bouckaert, R., and Heled, J. (2014) 'Densitree 2: Seeing Trees through the Forest', *BioRxiv*, 012401.

—— et al. (2012) 'Mapping the Origins and Expansion of the Indo-European Language Family', *Science*, 337/6097: 957–60.

Bouckaert, R. R., Bowern, C., and Atkinson, Q. D. (2018) 'The Origin and Expansion of Pama–Nyungan Languages across Australia', *Nature Ecology and Evolution*, 2: 741–9.

Bowern, C. (2018) 'Computational Phylogenetics', *Annual Review of Linguistics*, 4: 281–96.

—— (2016) 'Chirila: Contemporary and Historical Resources for the Indigenous Languages of Australia', *Language Documentation and Conservation*, 10.

Brinton, D. G. (1888) *The Language of Palæolithic Man*. Press of McCalla & Company.

Campbell, L. (1994) 'Putting Pronouns in Proper Perspective in Proposals for Remote Relationships among Native American Languages', in M. Langdon and L. Hinton (eds) *Proceedings of the Meeting of the Society for the Study of the Indigenous Languages of the Americas July 2-4, 1993, and the Hokan-Penutian Workshop, July 3, 1993*. Survey of California and Other Indian Languages. Berkeley, California: UC Berkeley Department of Linguistics.

—— (1997) 'Amerind Personal Pronouns: A Second Opinion', *Language*, 73/2: 339–51.

Casamiquela, R. M. (1983) *Nociones de gramática del gününa küne*. Paris: Centre National de la Recherche Scientifique.

Cerrón-Palomino, R. (1987) 'La Flexión de persona y número en el protoquechua', *INDIANA*, 11: 263–76.

—— (2000) *Lingüística aimara*. Cuzco: Centro de Estudios Regionales Andinos Bartolomé de Las Casas (CBC).

—— (2003) *Lingüística quechua*, 2nd edn. Cuzco: Centro de Estudios Regionales Andinos Bartolomé de Las Casas (CBC).

—— (2006) *El chipaya o la lengua de los hombres del agua*. Lima: Fondo Editorial, Pontificia Universidad Católica del Perú.

—— (2016) *El uro de la bahía de Puno*. Lima: Instituto Riva Agüero. Pontificia Universidad Católica del Perú.

Diamond, J., and Bellwood, P. (2003) 'Farmers and Their Languages: The First Expansions', *Science (New York, N.Y.)*, 300/5619: 597–603. https://doi.org/10.1126/science.1078208

Dixon, R. M. W. (1997) *The Rise and Fall of Languages*. Cambridge: Cambridge University Press.

Drummond, A. J. et al. (2006) 'Relaxed Phylogenetics and Dating with Confidence', *PLoS Biology*, 4/5: e88.

—— et al. (2012) 'Bayesian Phylogenetics with BEAUti and the BEAST 1.7', *Molecular Biology and Evolution*, 29/8: 1969–73.

Enfield, N. J. (2014) *Natural Causes of Language Frames, Biases and Cultural Transmission*. Berlin: Language Science Press.

Eriksen, L., and Danielsen, S. (2014) 'The Arawak Matrix', in O'Connor L. and Muysken P. C. (eds) *The Native Languages of South America: Origins, Development, Typology*, pp. 152–76. Cambridge: Cambridge University Press.

Estabrook, G. F., McMorris, F. R., and Meacham, C. A. (1985) 'Comparison of Undirected Phylogenetic Trees Based on Subtrees of Four Evolutionary Units', *Systematic Zoology*, 34/2: 193–200.

Gibson, M. L. (1996) *El Munichi: Un idioma que se extingue. Yarinacocha*, Pucallpa: Instituto Lingüístico de Verano.

Greenberg, J. H. (1960) 'General Classification of Central and South American Languages', in Wallace A. (ed.) *Men and Cultures: Fifth International Congress of Anthropological and Ethnological Sciences (1956)*, pp. 791–4. Philadelphia: University of Pennsylvania Press.

—— (1987) *Language in the Americas*. Stanford, California: Stanford University Press.

Grollemund, R. et al. (2015) 'Bantu Expansion Shows That Habitat Alters the Route and Pace of Human Dispersals', *Proceedings of the National Academy of Sciences*, 112/43: 13296–301.

Hammarström, H., Forkel, R., and Haspelmath, M. (2018) *Glottolog (Version 3.3)*. Jena: Max Planck Institute for the Science of Human History. Retrieved from http://glottolog.org

Haspelmath, M., and Tadmor, U. (eds). (2009) *WOLD*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved from http://wold.clld.org/.

Heckenberger, M. J. (2002) 'Rethinking the Arawakan Diaspora', in Hill J. D. and Santos-Granero F. (eds) *Comparative Arawakan Histories: Rethinking Language Family and Culture Area in Amazonia*, pp. 99–122. Urbana and Chicago: University of Illinois Press.

Jolkesky, M. (2016) *Estudo arqueo-ecolinguístico das terras tropicais sul-americanas* (Tese de Doutorado). Brasília, DF: Universidade de Brasília.

Kaufman, T. (1990) 'Language History in South America: What We Know and How to Know More', in Payne D. L. (ed.) *Amazonian Linguistics: Studies in Lowland South American Languages*, pp. 13–67. Austin: University of Texas Press.

—— (1994) 'The Native Languages of South America', in Mosley C. and Asher R. E. (eds) *Atlas of the World's Languages*, pp. 46–76. London: Routledge.

Key, M. R., and Comrie, B. (eds) (2015) *IDS*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved from http://ids.clld.org/.

Kroeber, A. L. (1913) 'The Determination of Linguistic Relationship', *Anthropos*, 8/2/3: 389–401.

List, J.-M., Greenhill, S., and Forkel, R. (2017) LingPy. A Python Library for Historical Linguistics (Version 2.6). Retrieved from http://lingpy.org.

—— (2019) 'Automatic Inference of Sound Correspondence Patterns across Multiple Languages', *Computational Linguistics*, 1/45: 137–61.

List, J. M., Greenhill, S. J., and Gray, R. D. (2017) 'The Potential of Automatic Word Comparison for Historical Linguistics', *PloS One*, 12/1: e0170046.

—— et al. (2018) 'Sequence Comparison in Computational Historical Linguistics', *Journal of Language Evolution*, 3/2: 130–44.

Macklin-Cordes, J. L., and Round, E. R. (2015) High-definition phonotactics reflect linguistic pasts. *Proceedings of the 6th Conference on Quantitative Investigations in Theoretical Linguistics*, 1–5.

Maddieson, I. (1984) *Patterns of Sounds*. Cambridge: Cambridge University Press.

Malvestitti, M., and Orden, M. E. (2014) *Günün a yajütshü: el Vocabulario Puelche documentado por Roberto Lehmann-Nitsche*. La Pampa: Universidad Nacional de la Pampa.

Matras, Y. (2007) 'The Borrowability of Structural Categories', in Matras Y. and Sakel J. (eds) *Grammatical Borrowing in Cross-Linguistic Perspective*, p. 44. Berlin & New York: Mouton de Gruyter.

—— (2009) *Language Contact*. Cambridge: Cambridge University Press.

Mufwene, S. S. (2005) 'Language Evolution: The Population Genetics Way', in Hauska G. (ed.) *Gene, Sprachen, und ihre Evolution*, pp. 30–52. Universitätsverlag Regensburg.

Muysken, P. C., and Hannss, K. (2006) 'Verbs in Uchumataqu', in *LOT Occasional Series*, pp. 215–33. Retrieved from http://dspace.library.uu.nl/handle/1874/296555.

——, and O'Connor, L. (eds). (2014) *The Native Languages of South America*. Cambridge: Cambridge University Press.

Nettle, D. (1999a) 'Linguistic Diversity of the Americas Can Be Reconciled with a Recent Colonization', *Proceedings of the National Academy of Sciences of the United States of America*, 96/6: 3325–9.

—— (1999b) *Linguistic Diversity*. Oxford: Oxford University Press.

Nichols, J., and Peterson, D. A. (1996) 'The Amerind Personal Pronouns', *Language*, 72/2: 336–71.

——, and —— (1998) 'Amerind Personal Pronouns: A Reply to Campbell', *Language*, 74/3: 605–14.

Overall, S. E. (2017) *A Grammar of Aguaruna*. Berlin, Boston: De Gruyter Mouton. Retrieved from https://www.degruyter.com/view/product/448714.

Pemberton, T. J., DeGiorgio, M., and Rosenberg, N. A. (2013) 'Population Structure in a Comprehensive Genomic Data Set on Human Microsatellite Variation', *G3: Genes, Genomes, Genetics*, 3/5: 891–907.

Peña, J. (2016) A Grammar of Wampis. Retrieved from http://scholarsbank.uoregon.edu/xmlui/handle/1794/19730.

Peyró García, M. (2005) 'Estructuras Gramaticales en el Glosario de la Lengua Atacameña (1896)', *LIAMES: Línguas Indígenas Americanas*, 5: 25–42.

Pompei, S., Loreto, V., and Tria, F. (2011) 'On the Accuracy of Language Trees', *Plos One*, 6/6: e20109.

Rambaut, A., and Drummond, A. J. (2012) *Tracer v1.5*.

——, and —— (2017) TreeAnnotator (Version 2.4.7). Retrieved from http://beast.community/treeannotator.

Reich, D. et al. (2012) 'Reconstructing Native American Population History', *Nature*, 488/7411: 370–4.

Rojas-Berscia, L. M. (2015) 'Mayna, the Lost Kawapanan Language', *LIAMES*, 15: 393–407.

—— (2019) 'From Kawapanan to Shawi: Topics in language variation and change', PhD dissertation. MPI Series 143. Nijmegen, The Netherlands, Radboud Universiteit Nijmegen, Max Planck Institute for Psycholinguistics.

——, and Nikulin, A. (2016) 'Nuevos alcances Para la reconstrucción léxica y fonológica del proto-cahuapana y más allá', in *Coloquio Amazónicas VI - Simposio de Fonología*. Leticia: Universidad Nacional de Colombia.

——, and Piepers, J. (in prep.). *The Valency Changing Operator -te: the Arawak Flux Hypothesis*.

Ruhlen, M. (1987) *A Guide to the World's Languages (Vol. 1: Classification)*. Stanford, California: Stanford University Press.

Saad, G. (2014) 'A Grammar Sketch of Shuar', MA thesis, Radboud Universiteit Nijmegen, Nijmegen.

Sandoval, J. R., The Genographic Project Consortium. et al. (2016) 'The Genetic History of Peruvian Quechua-Lamistas and Chankas: Uniparental DNA Patterns among Autochthonous Amazonian and Andean Populations', *Annals of Human Genetics*, 80/2: 88–101.

Sapir, E. (1929) 'The Status of Linguistics as a Science', *Language*, 207–14.

Sasse, H.-J. (2015) 'Syntactic Categories and Subcategories', in Kiss T. and Alexiadou A. (eds) *Syntax - Theory and Analysis. An International Handbook*, Vol. 3, pp. 158–217. Berlin: Mouton de Gruyter. Retrieved from https://www-degruyter-com.ezproxy.library.uq.edu.au/downloadpdf/books/9783110377408/9783110377408.158/9783110377408.158.pdf.

Suchard, M. A. et al. (2018) 'Bayesian Phylogenetic and Phylodynamic Data Integration Using BEAST 1.10', *Virus Evolution*, 4/1. Doi: 10.1093/ve/vey016.

Swadesh, M. (1954) 'Perspectives and Problems of Amerindian Comparative Linguistics', *WORD*, 10/2–3: 306–32.

Thomason, S. G., and Everett, D. L. (2001) 'Pronoun Borrowing', *Annual Meeting of the Berkeley Linguistics Society*, 27/1: 301.

Tornello, P. J. et al. (2011) *Introducción al Millcayac: idioma de los huarpes de Mendoza: textos de Luis de Valdivia*. Mendoza: Zeta Editores.

Valenzuela Bismarck, P. M. (2011) 'Contribuciones Para la reconstrucción del Proto-Cahuapana: comparación léxica y gramatical

de las lenguas jebero y chayahuita', in Adelaar W. F. H., Valenzuela P. M., and Zariquiey R. (eds) *Estudios en Lenguas Andinas y Amazónicas. Homenaje a Rodolfo Cerrón-Palomino*, pp. 274–304. Lima: Pontificia Universidad Católica del Perú.

Viegas Barros, J. P. (2017) Reconstrucción interna en pronombres e índices personales de la lengua Gününa Yajüch, pp. 1–14. *Presented at the Congreso Internacional de la Asociación de Lingüística y Filología de América Latina (ALFAL)*, Bogotá.

Willerman, R. (1994) 'The Phonetics of Pronouns: Articulatory Bases of Markedness', PhD thesis,. University of Texas at Austin.

Wundt, W. (1904) *Völkerpsychologie: Eine Untersuchung der Entwicklungsgesetze von Sprache, Mythus und Sitte*, Vol. 1. Leipzig: Verlag von Wilhelm Engelmann. Retrieved from https://ia802705.us.archive.org/28/items/vlkerpsycholog11wund/vlkerpsycholog11wund.pdf.

Zamponi, R. (2017) 'First-Person n and Second-Person m in Native America: A Fresh Look', *Italian Journal of Linguistics*, 29/2: 189–230.